Logistic_regression_preliminary analysis

Team: Who's there

Thursday, August 18, 2016

Introduction

The aim of the project was to be able to find a model that best predict the relationship between: * the number of people counted with the survey for a given class at a particular hour * Wi-fi log counted in that room at that hour

This will allow to see whether Wi-Fi log is a good predictor for estimating occupancy in a classroom.

Given the fact that the target feature was a percentage we tried to run a logistic regression to predict when the room was more likely to be empty or occupied.

Below we describes step by step all the analysis performed.

ANALYSIS

DATABASE CONNECTION AND DATASET

```
## Loading required package: DBI
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:DAAG':
##
##     ozone
```

First of all, we set up the connection to the database, using the following code:

```
connection <- dbConnect(MySQL(),user="root", password="",dbname="who_there_db", host="localhost")
```

Then we made a query to the database, in order to get all the groundth truth data collected in room B.002, B.004 and B.006 from 9 to 17 and the correspondent Wi-Fi Log measured in that time frame and rooms.

The dataset created had in total 216 rows and it will allow us to explore if Wi-Fi log can estimate precisely if a room is empty or occupied during a certain hour.

As a **target features** for the logistic regression we decided to create a binary categorical features with 2 levels using the percentage of the room full. In particular, we consider as empty all the cases in which the occupancy of the room was 0 and occupied when the occupancy of the room was higher than 0. Empty was indicated as 0, while Occupied as one.

As response variables or feature we considered Wi-Fi logs, which were summarised either as average of the logs counted for each room and for each hour or as maximum of the logs measured for each room and for each hour.

Together with the Wi-Fi log, we included in the data set the following features:

- **Date**, which we did not use in this analysis, because they just cover 2 weeks of Novemeber, but for future analyses they can be used to group observations by seasons or semesters or to finds seasonal trends for time series analyses.
- **Time**, which will be explored either as continous variable and as categorical to explore if the time of the day can have an affect on the Wi-Fi log. To do so we, bin the time in 4 ranges: early morning (9-11), late morning (11-13), early afternoon (13-15) and late afternoon (15-17). This will allow us to see if the Wi-Fi log accuracy was changing during the day. For example, it is more likely that all the electronic devices are fully powered early in the morning and consequently the Wi-fi log data can be more accurate or overestimating the occupancy of the room (i.e. more than one device per person). On the contrary in the afternoon, the devices may be more likely to be out of battery and it is possible that there are less devices in the room.
- **Module**, which we are not going to include in the analysis because the majority of the module present are for computer science, but for future analyses it will be possible to explore if the Wi-fi log accuracy in predicting the occupancy change acroos the courses. Science course or computer science course will more likely to use electronic devices during lecture than art students.
- **Course level**, which can indicate us whether electronic devices will be less used during different course level. For example, first and second level courses can be less practical and therefore laptop are not needed and that can decrease the number of devices connected. On the other hand, undergraduate might be more distracted during lecture and look at their phones during lectures. This will result in an increase of connection in that hour.
- **Tutorial**, which can affect the number of logged people. First of all, because tutorial divided the room in 2 and therefore there will be measured less people than expected.
- **Double_module**, categorical variable indicating wether in the class there are more than one module, increasing the number of people expected in the room.
- **Double_module**, categorical variable indicating wether in the class went ahead to check for false positive.

The resulting data set is printed below:

```
head(AnalysisTable)
```

```
##   Room       Date Time    Module Course_Level Tutorial Double_module
## 1    1 2015-11-03    9         0            0        0             0
## 2    1 2015-11-04    9 COMP30190            3        0             0
## 3    1 2015-11-05    9         0            0        0             0
## 4    1 2015-11-06    9 COMP30220            3        0             0
## 5    1 2015-11-09    9 COMP30190            3        0             0
## 6    1 2015-11-10    9         0            0        0             0
##   Class_went_ahead Capacity Percentage_room_full Wifi_Average_clients
```

```
## 1                1       90                 0.00              4.7500
## 2                1       90                 0.25             13.4545
## 3                1       90                 0.00              6.8333
## 4                1       90                 0.00              2.4167
## 5                1       90                 0.25             14.7273
## 6                1       90                 0.00              2.2727
##   Wifi_Max_clients Binary_Occupancy   Factor_Time
## 1               21                0 Early Morning
## 2               15                1 Early Morning
## 3               29                0 Early Morning
## 4                3                0 Early Morning
## 5               18                1 Early Morning
## 6               14                0 Early Morning
```

DATA QUALITY REPORT

Before running any analyses, we carried out the data quality report to check for any issue related to the variable (e.g. outlier, skewed distribution, NaN values) and solutions we will implement to solve them.

Initially set all the categorical variables as factors and then we printed the descriptive statistic for all the features.

```
summary(AnalysisTable)
```

```
##       Room       Date                Time          Module
##  Min.   :1   Length:216         Min.   : 9.00   Length:216
##  1st Qu.:1   Class :character   1st Qu.:10.75   Class :character
##  Median :2   Mode  :character   Median :12.50   Mode  :character
##  Mean   :2                      Mean   :12.50
##  3rd Qu.:3                      3rd Qu.:14.25
##  Max.   :3                      Max.   :16.00
##  Course_Level Tutorial Double_module Class_went_ahead    Capacity
##  0:59         0:210    0:210         0: 22            Min.   : 90.0
##  1:14         1:  6    1:  6         1:194            1st Qu.: 90.0
##  2:23                                                 Median : 90.0
##  3:76                                                 Mean   :113.3
##  4:40                                                 3rd Qu.:160.0
##  5: 4                                                 Max.   :160.0
##  Percentage_room_full Wifi_Average_clients Wifi_Max_clients
##  Min.   :0.00         Min.   :  0.00       Min.   :  0.00
##  1st Qu.:0.00         1st Qu.: 11.61       1st Qu.: 18.75
##  Median :0.25         Median : 23.67       Median : 32.50
##  Mean   :0.25         Mean   : 30.33       Mean   : 40.01
##  3rd Qu.:0.25         3rd Qu.: 43.88       3rd Qu.: 55.25
##  Max.   :1.00         Max.   :192.92       Max.   :230.00
##  Binary_Occupancy          Factor_Time
##  0: 67            Early Morning  :54
##  1:149            Late Morning   :54
##                   Early Afternoon:54
##                   Late Afternoon :54
##
##
```

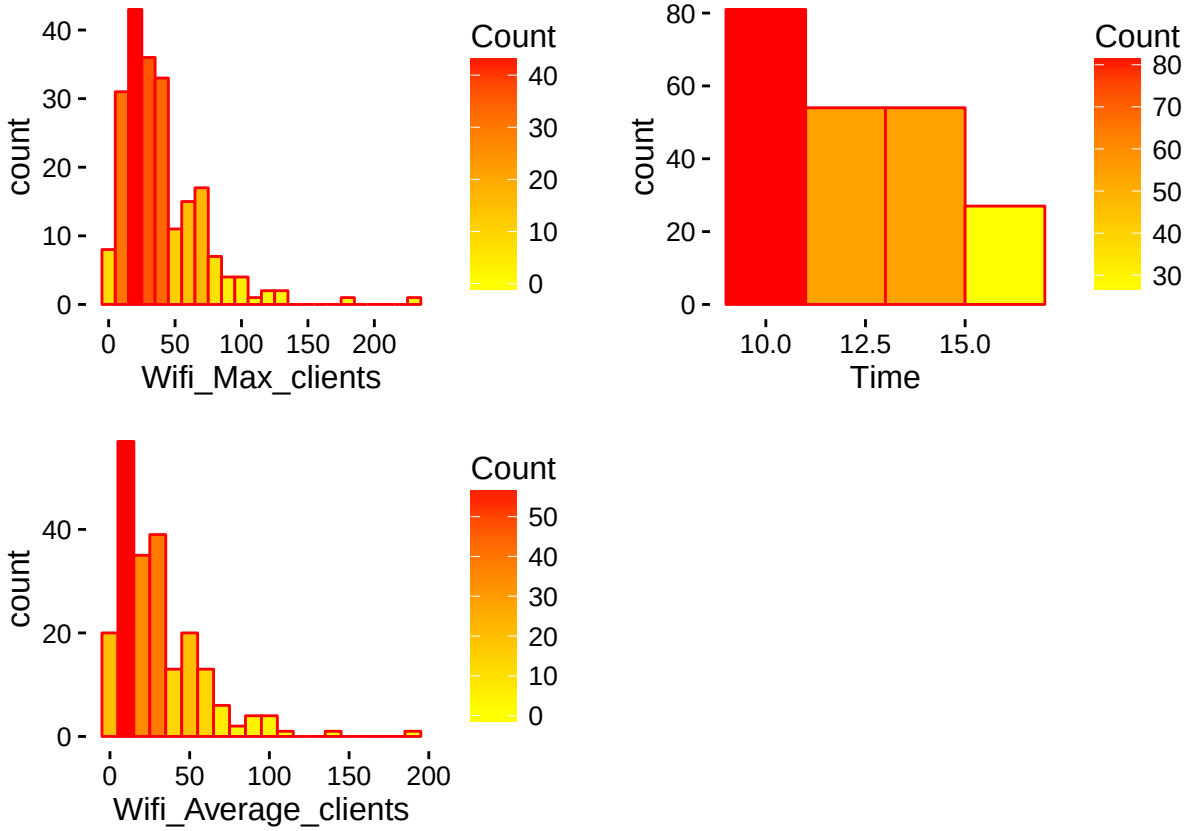
From this we could see that NaN values were not present in the data set. We could notice that the observations for the features Tutorials and Double_model were not even distributed across the 2 levels of the variables. In fact, only 6 observations were present for tutorial class and for double module class. Therefore, we decided to discard both the features, because they will be not informative for the analysis. Similarly for the feature class_went_ahead the majority of the lectures did occur and we decided to discard it. Furthermore for the variables Wifi_Average_clients and Wifi_Max_clients, it seems that there are few outliers, since the median is lower than the mean and the max values are far higher than the mean values. We will going to explore this issues with histogram and boxplots.

Exploratory graphs

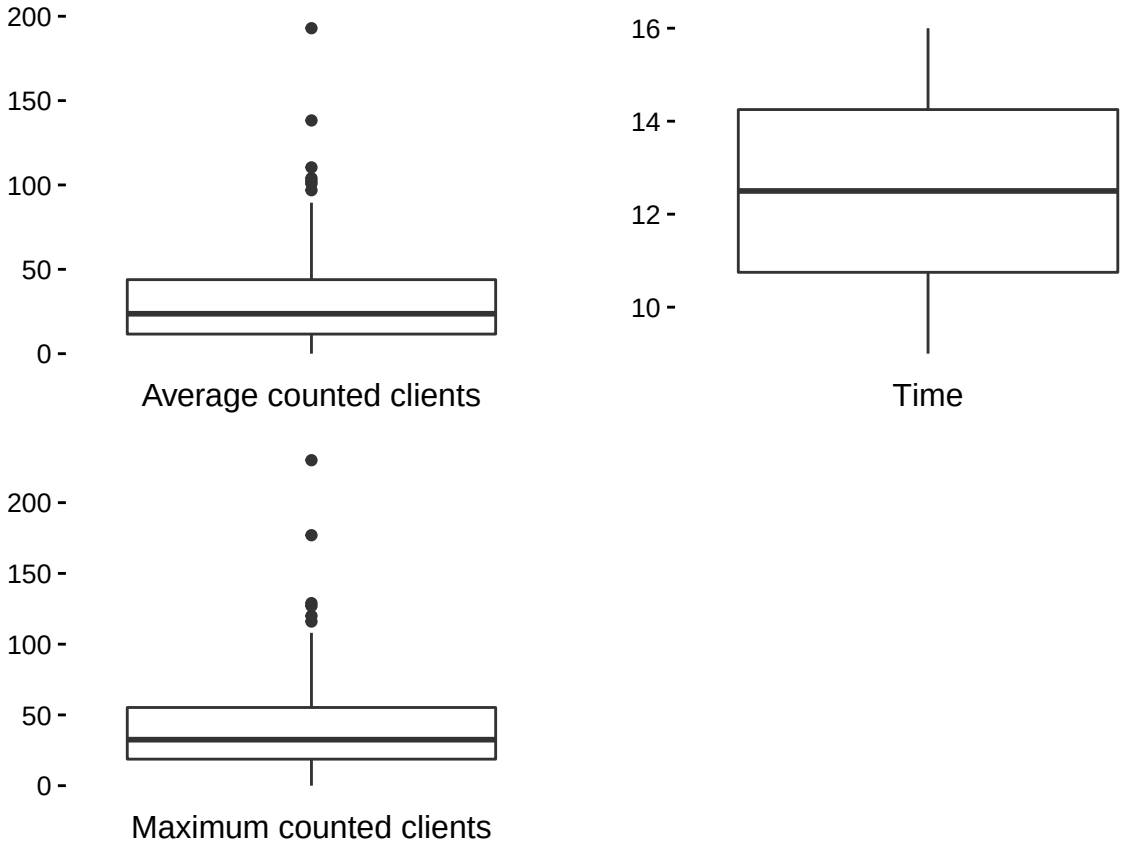
For exploring possible issues related with the continuous variables we plotted histograms and boxplots.

Histograms

```
## Loading required package: grid
```



Form the histograms we could see that the distribution of the feature Wifi Maximum_client (i.e. the Maximum number of devices logged in one hour lecture) was skewed to the left, indicating that the in the majority of the lecture were counted no more than 40 people. Furthermore, we could see that there are potential outliers (values > 150). Similar pattern was observed for the feature Wifi_Average_clients. Feature times had as well a skewed distribution, suggesting that the majority of the lectures were concentrating during the early morning and they were decreasing towards the afternoon. ### Box plots.



From the boxplots, all the trends observed in the histograms were confirmed.

For categorical variables we plot bar plot graphs.

Bar plots.

From the barplots, we could see that observations were equally distributed across all the levels of the feature Room and Factor Time. On the contrary, there were more observations for non lectures and level 3 courses and classes was observed mostly occupied. No issues were detected for those features.

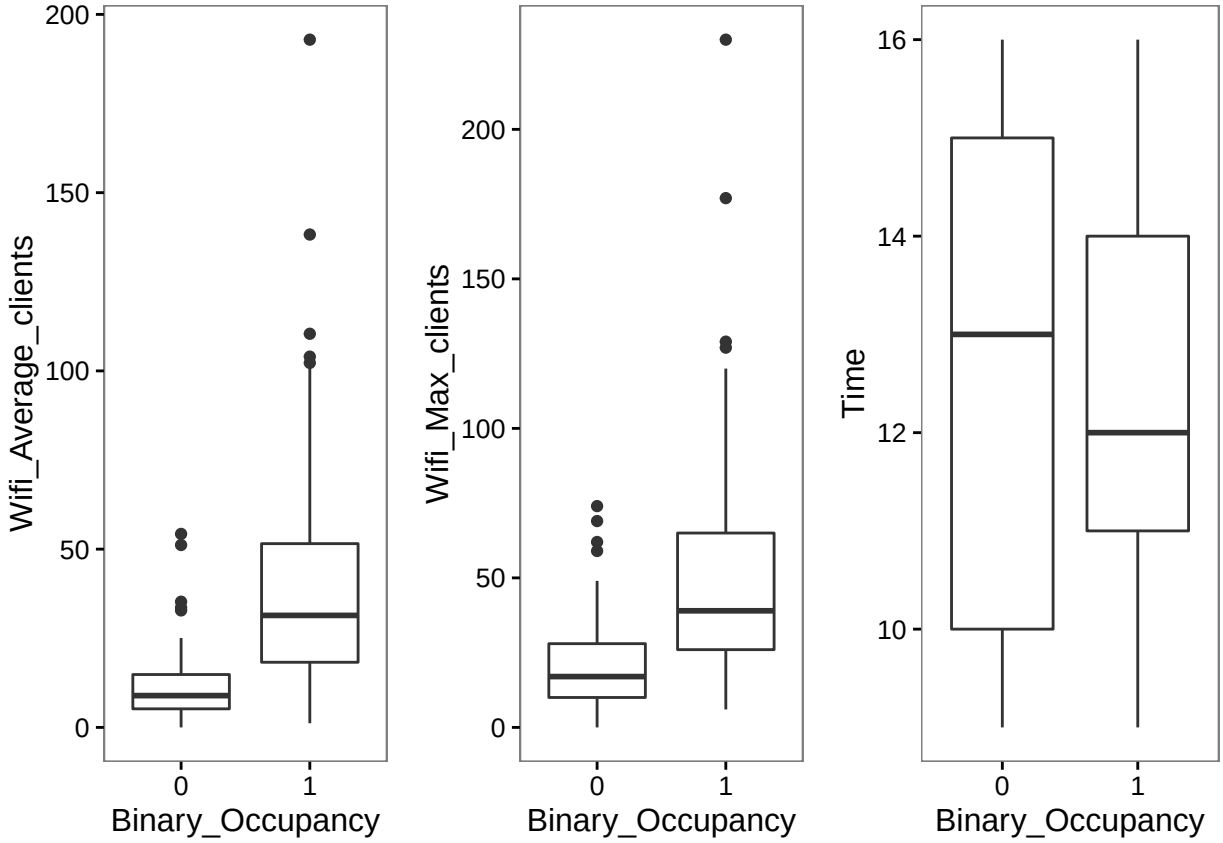
Summary.

Features	Issues	Planned Solution
Room	None	None
Time	Distibution skewed to the left	To solve during analysis
Factor Time	None	None
Course level	None	None
Tutorial	Uneven representation of the level	Discarded from the analysis
Double Module	Uneven representation of the level	Discarded from the analysis
Class went ahead	Uneven representation of the level	Discarded from the analysis
Wifi Average clients	Distibution skewed to the left & outliers	To solve during analysis
Wifi Maximum clients	Distibution skewed to the left & outliers	To solve during analysis
Binary Occupancy	None	None

FEATURES AFFECTING THE TARGET FEATURE

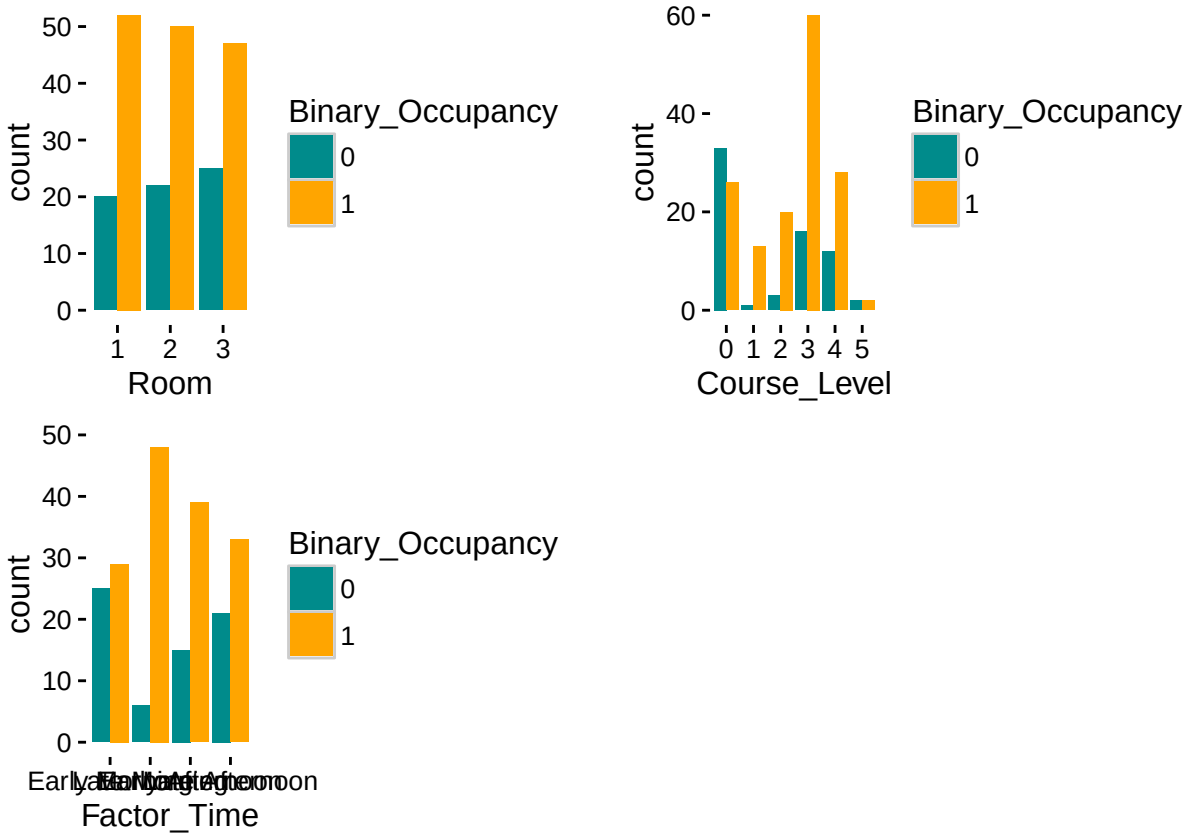
The next step of the analysis was to see which feature really affect the target feature for deciding which features we would include into the model.

For the continuous features we used box plots, while for categorical we use bar plots.



From the boxplots we could see that when the room was indicated as empty either the average and the maximum Wi-Fi logs were close to zero. On the contrary, when the room was occupied the average and the maximum Wi-Fi logs were different from zero. The difference between the 2 levels will probably be higher without the outliers. Therefore, we can conclude that either average Binary_Occupancy and maximum wifi counted clients are a good predictor of the binary occupancy and we are going to run 2 models: one for exploring the relationship between Binary_Occupancy and Average counted clients and another for Survey counted clients and Maximum counted clients.

Barplots

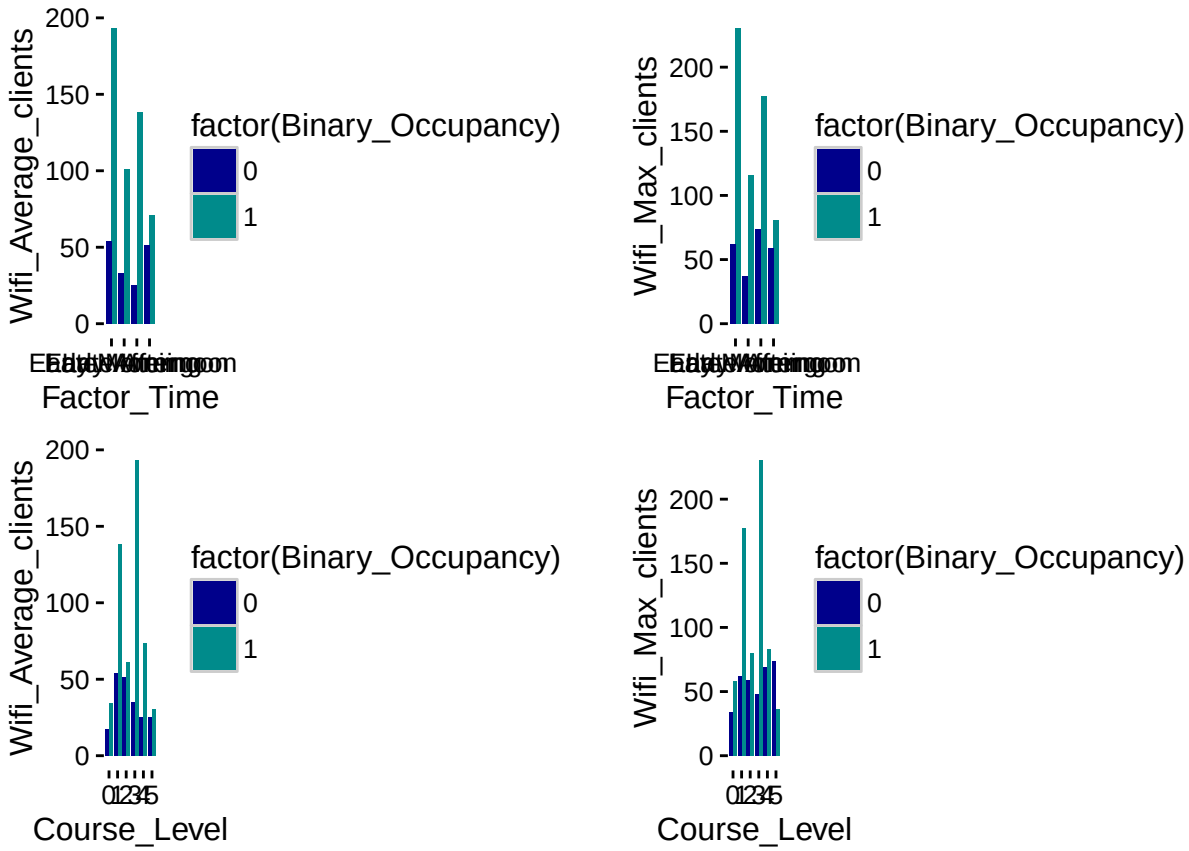


From the barplots, we could see that the occupancy was not affected by the time of room. On the contrary, the occupancy of the room increase for level3 and 4 courses, while was lower for level1, 2 and 5, which in our case represents career talk or CS meeting. The occupancy of the room in late monrning and late afternoon seemed to be higher. Therefore, in our model we are going to explore whether course level and time of the day affect the chance of a room to be empty or occupied.

Interacting effect on the target feature

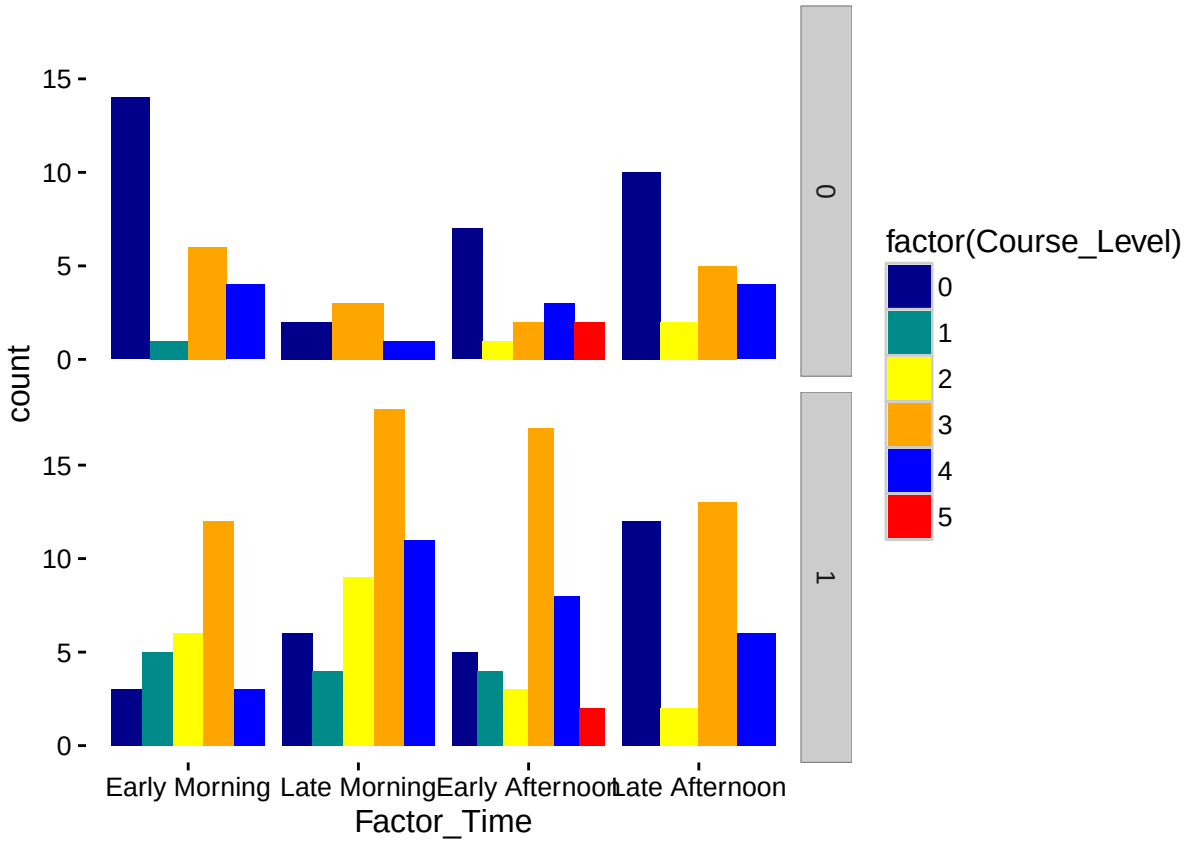
The last step before the regression was to explore the interacting effect of the features on the target features.

First of all we explore the interactive effect between Average conted clients with all the other features using bar plots.



From the graphs we could see that the average and the maximum number of devices counted with the Wifi-logs were lower in rooms indicated as empty by the survey across all the level of the time factor, indicating that there was not an interaction between Wifi average count and time and between Wifi maximum count and time. Similar pattern was found across all the levels of course level, suggesting that there was not interaction between Wifi average count and course levels and between Wifi maximum count and course levels.

For exploring the interaction between Time and course level we did the following bar plot:



From this graph we cannot detect any clear pattern, therefore, we are not going to explore the interactive effect between Time and course level.

Consequently we are going to explore the following 2 models: * $Binary_Occupancy\ AverageWifioccupancy + Time + Course_Level$ * $Binary_Occupancy\ MaximumWifioccupancy + Time + Course_Level$

Analysis

For the preliminary analysis for time constraint we run the analysis using the Validation set approach, which consists in dividing the dataset in a training and a test approach. Since the dataset was not that big, we decided to divide it in 60% for training and 40% for testing. This will give to the test dataset enough data for running the linear model. We are aware of the limitation of the Validation Set Approach and in the next analyses we are going to run the model with a 10-fold cross validation.

Models	Accuracy
Binary_Occupancy ~Wifi_Max_clients + Room + Factor_Time + Course_Level	0.7816092
Binary_Occupancy ~Wifi_Average_clients + Room + Factor_Time + Course_Level	0.8275862

The second model was the one with the best accuracy, so we decided to apply it to the whole data set and we explore its residuals to see if it was a good model. The graph plotting the residuals, against predicted values and the logistic regression in black with green showing the confidence interval showed that the logistic regression is quite close to the dotted line as should be expected and it shows the presence of potential outliers.

For the next analysis we will try to implement a multinomial regression with a k-fold cross validation and we

might explore if this is better without the outliers.